

MODELING OF HEAD RELATED TRANSFER  
FUNCTIONS FOR IMMERSIVE AUDIO USING  
A STATE-SPACE APPROACH

5

This application claims the benefit of U.S. provisional application Serial No. 60/238,872, filed October 6, 2000.

BACKGROUND OF THE INVENTION

10

15

20

25

30

35

40

Applications for 3-D sound rendering include teleimmersion; augmented and virtual reality for manufacturing and entertainment; teleconferencing and telepresence; air-traffic control; pilot warning and guidance systems; displays for the visually impaired; distance learning; and professional sound and picturing editing for television and film. Work on sound localization finds its roots as early as the beginning of the twentieth century when Lord Rayleigh first presented the Duplex Theory that emphasized the importance of Interaural Time Differences (ITD) and Interaural Amplitude Differences (IAD) in source localization. It is notable that human listeners can detect ITD's as small as 7  $\mu$ s which makes it an important cue for localization. Nevertheless, ITD's and IAD's alone are not sufficient to explain localization of sounds in the median plane, in which ITD's and IAD's are both zero.

Variations in the spectrum as a function of azimuth and elevation angles also play a key role in sound localization. These variations arise mainly from reflection and diffraction effects caused by the outer ear (pinna) that give rise to amplitude and phase changes for each angle. These effects are described by a set of functions known as the Head-Related Transfer Functions (HRTF's).

One of the key drawbacks of 3-D audio rendering systems arise from the fact that each listener has HRTF's that are unique for each angle. Measurement of HRTF's is a tedious process that is impractical to perform for every possible angle around the listener. Typically, a relatively small number of angles are measured and various methods are used to generate the HRTF's for an arbitrary angle. Previous work in this area includes modeling using principal component analysis, as well as spatial feature extraction and regulation.

Disclosed is a two-layer method of modeling HRTF's for immersive audio rendering systems. This method allows for two degrees of control over the accuracy of the model. For example, increasing the number of measured HRTF's improves the spatial resolution of the system. On the other hand, increasing the order of the model extracted from each measured HRTF improves the accuracy of the response for each measured direction. Kung's method was used to convert the time-domain representation of HRTF's in state-space form. The models were compared both in their Finite Impulse Response (FIR) filter form and their state-space form. It is clear that the state-space method can achieve greater accuracy with lower order filters. This was also shown using a balanced model truncation method. Although an Infinite Impulse Response (IIR) equivalent of the state-space filter could be used without any theoretical loss of accuracy, it can often lead to numerical errors causing an unstable system, due to the large number of poles in the filter. State-space filters do not suffer as much from the instability problems of IIR filters, but require a larger number of parameters for a filter of the same order. However, considering that there are similarities among the impulse responses for different azimuths and elevations, a combined single system model for all directions can provide, as we will show, a significant reduction.

Previous work on HRTF modeling has mainly focused on methods that attempt to model each direction-specific transformation as a separate transfer function. In this paper we present a method that attempts to provide a single model for the entire 3-D space. The model builds on a generalization of work by Haneda et al, in which the authors proposed a model that shares common poles (but not zeros) for all directions.

1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035	2036	2037	2038	2039	2040	2041	2042	2043	2044	2045	2046	2047	2048	2049	2050	2051	2052	2053	2054	2055	2056	2057	2058	2059	2060	2061	2062	2063	2064	2065	2066	2067	2068	2069	2070	2071	2072	2073	2074	2075	2076	2077	2078	2079	2080	2081	2082	2083	2084	2085	2086	2087	2088	2089	2090	2091	2092	2093	2094	2095	2096	2097	2098	2099	2100	2101	2102	2103	2104	2105	2106	2107	2108	2109	2110	2111	2112	2113	2114	2115	2116	2117	2118	2119	2120	2121	2122	2123	2124	2125	2126	2127	2128	2129	2130	2131	2132	2133	2134	2135	2136	2137	2138	2139	2140	2141	2142	2143	2144	2145	2146	2147	2148	2149	2150	2151	2152	2153	2154	2155	2156	2157	2158	2159	2160	2161	2162	2163	2164	2165	2166	2167	2168	2169	2170	2171	2172	2173	2174	2175	2176	2177	2178	2179	2180	2181	2182	2183	2184	2185	2186	2187	2188	2189	2190	2191	2192	2193	2194	2195	2196	2197	2198	2199	2200	2201	2202	2203	2204	2205	2206	2207	2208	2209	2210	2211	2212	2213	2214	2215	2216	2217	2218	2219	2220	2221	2222	2223	2224	2225	2226	2227	2228	2229	2230	2231	2232	2233	2234	2235	2236	2237	2238	2239	2240	2241	2242	2243	2244	2245	2246	2247	2248	2249	2250	2251	2252	2253	2254	2255	2256	2257	2258	2259	2260	2261	2262	2263	2264	2265	2266	2267	2268	2269	2270	2271	2272	2273	2274	2275	2276	2277	2278	2279	2280	2281	2282	2283	2284	2285	2286	2287	2288	2289	2290	2291	2292	2293	2294	2295	2296	2297	2298	2299	2300	2301	2302	2303	2304	2305	2306	2307	2308	2309	2310	2311	2312	2313	2314	2315	2316	2317	2318	2319	2320	2321	2322	2323	2324	2325	2326	2327	2328	2329	2330	2331	2332	2333	2334	2335	2336	2337	2338	2339	2340	2341	2342	2343	2344	2345	2346	2347	2348	2349	2350	2351	2352	2353	2354	2355	2356	2357	2358	2359	2360	2361	2362	2363	2364	2365	2366	2367	2368	2369	2370	2371	2372	2373	2374	2375	2376	2377	2378	2379	2380	2381	2382	2383	2384	2385	2386	2387	2388	2389	2390	2391	2392	2393	2394	2395	2396	2397	2398	2399	2400	2401	2402	2403	2404	2405	2406	2407	2408	2409	2410	2411	2412	2413	2414	2415	2416	2417	2418	2419	2420	2421	2422	2423	2424	2425	2426	2427	2428	2429
------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------

5

BRIEF DESCRIPTION OF THE DRAWINGS

FIGURE 1 is a flow diagram showing how the unprocessed signals are passed to the algorithm along with the desired azimuth and elevation angles of projection;

FIGURE 2 is a graphical representation of the delay in samples versus the angle measured related to the ear;

FIGURE 3 is a depiction of proposed convention of measuring azimuth in order to have a single delay and gain function for both ears;

FIGURE 4 is a graphical representation of the energy is signal versus the angle measured relative to the ear;

FIGURE 5 is a frequency domain of measured and simulated impulse responses for a model created with a  $30^\circ$  resolution.  $\theta = 40$ , and  $\theta = 50$  were not used for the creation of the model;

FIGURE 6 is a detail of the time domain of Figure 5;

FIGURE 7 is a model used is reduced down to 191 states from an original size of 600 states. Accuracy has not decreased significantly ; and

FIGURE 8 is 12 models of total 192 states. Accuracy has dropped significantly in comparison with Figure 7 although model size is the same.

# DETAILED DESCRIPTION OF THE INVENTION

One way to spatially render 3-D sound is to filter a monaural (non-directional) signal with the HRTF's of the desired direction. This involves a single filter per ear for each direction and a selection of the correct filter taps through a lookup table. The main disadvantage of this process is that only one direction can be rendered at a time and interpolation can be problematic. In our method, we extract the important cues of ITD and IAD as a separate layer, thus avoiding the problem of dual half-impulse responses created by interpolation. The second layer of the interpolation deals with the angle-dependent spectrum variations (Figure 1). This is a multiple-input single-output system (for each channel) which we created in state-space form.

The signal for any angle  $\theta$  can be fed to the input corresponding to that angle, or if there is no input corresponding to  $\theta$  then the signal can be split into the two adjacent inputs (or more in the case of both azimuth and elevation variations). In order to proceed with the two-layered model described above, we first extract the delay from the measured impulse responses. Figure 2 shows the delay extracted from the measurements and fitted with a sixth order polynomial.

It should be noted that here the azimuth is measured from the center of the head relative to the midcoronal and towards the face as shown in Figure 3 and not relative to the midsagittal and clockwise as a common practice. For example, the azimuth of  $270^\circ$  relative to the midsagittal corresponds to  $180^\circ$  for the right ear but to  $0^\circ$  for the left ear measured with this proposed convention. This method of representation was chosen because it allows us to use a common delay function for both ears.

Similarly, we can approximate the gain with a 14th order polynomial as in figure 4. The advantages of polynomial fitting are not so obvious when only one elevation is considered, but become more evident when the entire 3-D space is taken into consideration.

The measurements used in this paper include impulse responses taken using a KEMAR dummy head. These 512-point impulse responses can be used as an FIR model against which our comparisons will be based. A one input-one output case is briefly described below.

Consider an impulse response model of a causal, stable, multivariable and linear time-invariant system. If the system state space model is

$$x(n+1) = Ax(n) + Bu(n)$$

$$y(n) = Cx(n) + Du(n)$$

and an impulse is applied to the system then (assuming that  $u_0 = 1$ , without loss of generality):

$$\begin{aligned} y_0 &= D \\ x_1 &= B & y_1 &= CB \\ x_2 &= AB & y_2 &= CAB \\ x_3 &= A^2B & y_3 &= CA^2B \\ &\dots & &\dots \\ &\dots & &\dots \\ x_N &= A^NB & y_N &= CA^NB \end{aligned}$$

Forming the above into a matrix:

$$\begin{bmatrix} y(n) \\ y(n+1) \\ y(n+2) \\ \vdots \end{bmatrix} = \begin{bmatrix} CB & CAB & CA^2B & \dots \\ CAB & CA^2B & CA^3B & \dots \\ CA^2B & CA^3B & \dots & \dots \\ CA^3B & CA^4B & \dots & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} u(n) \\ 0 \\ \vdots \\ \vdots \end{bmatrix}$$

Separating the Handel matrix (i.e., the matrix that in position (i, j) is  $CA^{1+j-1}B$ ) and expressing it in its Singular Value Decomposition (SVD) components:

$$H = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \end{bmatrix} \cdot \begin{bmatrix} B & AB & A^2B & \dots \end{bmatrix} = WG = USV^T$$

where U, V are unitary matrices and  $\Sigma$  contains the singular values along its diagonal in decreasing magnitude, i.e.,

$$\Sigma = \text{Diag}[\sigma_1, \sigma_2, \sigma_3, \dots, \sigma_r, \sigma_{r+1}, \dots, \sigma_{N+1}]$$

and  $\Omega$  and  $\Gamma$  are the extended observability and reachability matrices that can be expressed in terms of the SVD components of H as:

$$W = US^{\frac{1}{2}} \text{ and } G = S^{\frac{1}{2}}V^T$$

One way to reduce the model is to use

$$H = \begin{bmatrix} U_n & \overline{U_n} \end{bmatrix} \cdot \begin{bmatrix} S_n & 0 \\ 0 & \overline{S_n} \end{bmatrix} \begin{bmatrix} V_n \\ \dots T \\ V_n \end{bmatrix}$$

and reduce  $\Omega$  to  $\Gamma$  to:

$$W_n = U_n S_n^{\frac{1}{2}} \text{ and } G_n = S_n^{\frac{1}{2}} V_n^T$$

5 This will give:

$$\begin{aligned} A &= S^{-\frac{1}{2}} U_n^T U_n^{\uparrow} S^{\frac{1}{2}} & C &= U_n^{\uparrow} S^{\frac{1}{2}} \\ B &= S^{\frac{1}{2}} (V_n^{\uparrow})^T & D &= y_0 \end{aligned}$$

15 While there are several definitions for  $U_n$  and  $U_n^{\uparrow}$ , one that also guarantees stability is

$$U_n = \begin{bmatrix} U_n^1 \\ \vdots \\ U_n^{N-1} \\ U_n^N \end{bmatrix} \text{ and } U_n^{\uparrow} = \begin{bmatrix} U_n^2 \\ \vdots \\ U_n^N \\ O \end{bmatrix}$$

25 To achieve higher speeds in model creation and the ability to handle any model size. The method is performed on each impulse response separately. This avoids the dimension increase of the Hankel matrix and consequently drops the computational cost of the SVD significantly since SVD is an  $O(3)$  operation. The individual state-space models are combined in a single model to form the final model. Further reduction can be achieved on the resulting model if desired.

35 The advantages of the two-layer HRTF model can better be observed by examining a few representative impulse responses. Figures 5 and 6 show the measured data with a dashed line and the

simulated data with a solid line. The model was created with data measured every 30°, and therefore only data from the first and last plot of each figure were used for the creation of the model. The other two simulated responses in the plot correspond to data synthesized from the 30° and 60° inputs of the state-space model. For example, angle 40° corresponds to  $\frac{2}{3}$  of the input signal being fed through the 30° input, while the remaining  $\frac{1}{3}$  is input to the 60° direction. As expected, the two main cues of delay and gain were preserved in the impulse response since they are generated from a separate, very accurate layer. The second layer can then be reduced according to the desired accuracy.

Figure 7 shows the performance of a further reduced state space model. The model was reduced to less than a third its initial size (down to 191 states from 600). As can be seen from the figures, there was some minor loss of accuracy. Figure 8 displays the performance of an equivalent model size that was created by reducing each individual HRTF to a 16 state model. These models correspond to a combined model of 192 states that is of equivalent size to the previous combined model but that performs very poorly. The advantage of performing the reduction to the combined invention is clearly evident.

Although the state-space model is computationally expensive compared to an FIR filter, it provides several advantages over the latter while avoiding some of the disadvantages of IIR filters. Recent advances in FPGA technology allow large matrix multiplications at very high speeds that would make construction of a larger size state-space device possible. Others consider  $N \times N$  times  $N \times N$  matrix multiplication, which can be extended to  $N \times N$  times  $N \times 1$  multiplication (the most expensive operation in the state space representation).  $N$  can be given by:

$$\frac{N^2}{p \times f_{FPGA}} < \frac{1}{44.1kHz}$$

for a signal sampled at 44.1kHz, where  $f_{FPGA}$  is the FPGA clock frequency and  $p$  is the number of parallel multipliers.

Today's FPGA's with speeds exceeding 150MHz and  $p > 100$  can easily handle state-space models of more than 500 states built on a single FPGA. As technology in this field is advancing with the System On a Chip model rapidly gaining ground, it will not be long before state-space models of more than a thousand states can be calculated in real time.

Another advantage that comes with the use of a state-space device is memory, which eliminates the audible "clicking" noise heard when changing from filter to filter. In fact, a model with many states eliminates the need for interpolation due to the memory. Interpolation, by passing a signal to two inputs at once, is however desirable to avoid sudden jumps in space of the virtual source.

Finally, we have demonstrated that while a single model for the whole space can achieve spatial rendering of multiple sources at once, it can also result in a smaller size than the individual models for all directions combined.